

2018/09/25-2018/09/30周报

DONE

1. 阅读《中国计算机学会通讯》

总体而言，通讯里面的文章涉及到硬件的部分比较难懂，或是比较浅，或是作者假设了读者懂这些概念，写的比较简略，阅读起来有些半知半解。

- 《面向图计算的加速器》是关于硬件方面的研究，主要分为三种，FPGA的加速器目标是加速访问存储，ASIC加速器主要基于并行的设计，新型器件加速器则主要是降低数据通讯的开销。最后作者提到，图加速器对开发人员熟悉硬件的程度要求非常高，C/C++是较好的算法语言。
- 《大图数据分析系统综述》对图分析需求进行了综述：主要分成四大类常见问题：1. 图搜索。2. 社团检测。3. 节点重要性和相关性分析。4. 图匹配查询。之后则总结了一些图分析系统：
 - 低层次抽象的分析系统：主要将一些图计算的基本操作抽象成接口，用户可以调用这些接口进行编程。比如Google的pregel分布式数据管理系统，主要的理念是主节点对图进行划分，子节点承担各个部分的计算任务。微软的Tinnity则是基于内存的分布式图数据管理系统。GraphX则是将图计算任务分成两类：图并行计算任务（节点层次上的计算），数据并行计算任务（整个图层次上的计算）
 - 高层次抽象的分析系统：用户只需要输入描述性的查询分析语句来进行分析。比如Neo4j/EmptyHeaded，这些系统一般都讲一些常见的算法进行整合，并提供一些描述性的查询分析语句来用于执行。
 - 基于RDF的知识图谱数据管理分析系统：主要介绍了北大研发的gStore，采用了图计算的方法来试图解决知识图谱的问题。
- 《简析面向图计算的运行时系统》主要考虑的问题是图计算过程中的负载均衡问题，涉及到一些内存分配/流水处理/IO方面的问题。
- 《分布式图计算》一文总结了一些图计算的特点（挑战）：

1. 高访存计算比，大量开销在IO上。
2. 数据局部性差，在利用处理器cache的时候，由于图数据的局部性差，无法进行复用而导致开销较高。
3. 类型与操作多样性：图的种类多，对图进行的操作也多。
4. 规模巨大，结构不规则，以及幂律分布的存在。

在这之后，作者总结了一个图计算的一般模型：首先进行数据划分（基于点划分/基于边划分/混合划分），之后通过一个模型（框架）或者直接进行数据传输，将图数据传递给任务执行。进行任务执行的系统模型，又可以分成三类：1. 同步模型，每一步执行任务，都需要进行同步，这样就导致了开销较大。2. 异步模型，节点单独进行任务执行，不需要同步，速度快。3. 数据流模型：由数据流驱动的模式。

- 《图计算在阿里巴巴的应用与挑战》，首先提出了阿里遇到的挑战：
 1. 分布式图遍历：利用分布式计算进行存储，并且加速对图数据的遍历。
 2. 渐进式计算，主要思想是先返回一个粗略结果，同时去做更精确的计算。
 3. 大图可视化，有限空间对大规模图进行可视化。

图计算对阿里的意义则是进行模式匹配，最主要的是对一些欺诈/作弊的行为进行检测，比如“刷单”这种普遍存在的作弊行为。阿里对模式匹配进行优化主要由四种方法：

1. 剪枝/并行
 2. 建立索引
 3. 增量计算：也就是复用以前的计算结果
 4. 硬件加速
2. vis报告的PPT制作。因为之前一直在赶chi，郭博也在赶毕业论文，所以抓紧时间在这两天做了一个比较粗糙的版本。国庆回来之后会抓紧时间优化。

工作时长

每周基本固定九点半前到（周四上午较晚才到），晚上十一点左右回寝室。

计划

短期计划（下周）

在国庆的路上阅读一些今年vis的相关图可视化paper。

中期计划（十月）

1. vis会议
2. paper阅读

长期计划（本学期）

1. 继续做大图可视化引擎，想以此为契机锻炼自己的代码能力，并将图可视化能够作为组件存在方便大家使用。
2. 继续巩固自己的前端基础。
3. 了解更多机器学习、数据挖掘相关的算法。